Large-scale empirical identification of candidate comparators for pharmacoepidemiological studies

Justin Bohn, ScD
Global Epidemiology Organization
Johnson & Johnson

Research Spotlight
OHDSI Community Call
22 July 2025

See the paper in
Drug Safety

Johnson&Johnson
Innovative Medicine

The traditional approach to comparator selection

New user (NU) cohort design has emerged as pharmacoepidemiological best practice

- But requires selection of a comparator appropriate to the target-outcome relationship under study

May be *several* known alternatives to the target → partially subjective

- What have other people done?
- Which alternatives are available in my preferred data source?
- Which gives me the largest sample?
- Which do I think will be least confounded?
- Which will give me the effect estimate I want? ☠

The actual comparison of interest may be to a non-user group → challenging

- "Negative controls exposures"
- "Null comparators"

Note that the third situation is often implicit in safety evaluations, even when investigators know better than to say so

Can we provide the investigator with empirical information to help?

Motivation

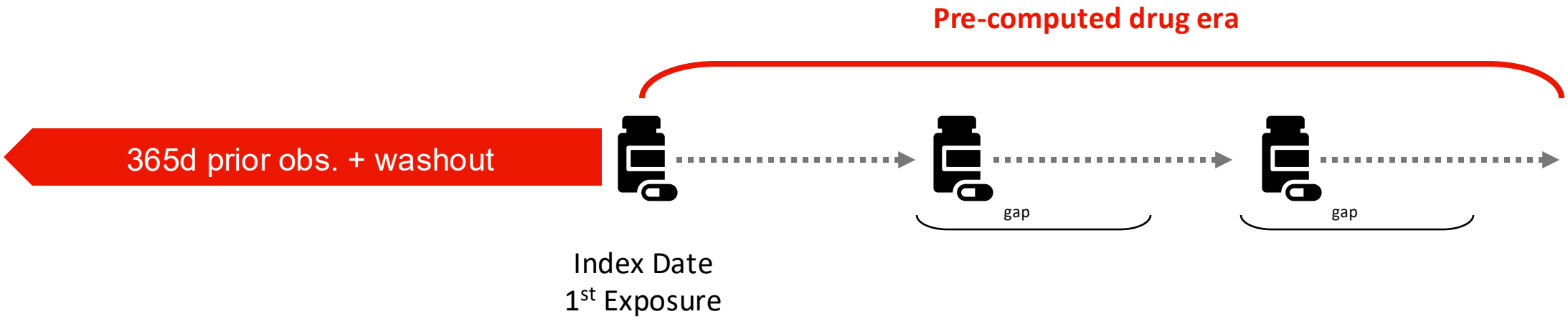
Our proposed solution

- Build a tool that identifies new-user cohorts with similar covariate profiles
  - Define new user cohorts
  - Extract covariate data
  - Compute similarity metrics
  - Rank comparators
  - Explore comparisons
- Evaluate the performance of the tool
  - Do highly ranked comparators appear clinically relevant?
  - Does the computed similarity align with known relationships between drugs (e.g., in terms of anatomical, therapeutic, chemical groupings?)
  - Does the computed similarity correspond to more familiar measures of covariate balance like standardized differences?
  - Does the tool rank highly those comparators chosen by high-quality publications?

Methods

Building a cohort similarity pipeline

Generating new-user cohorts for thousands of drugs and classes



Repeat for every RxNorm ingredient (~1,500)

Repeat for every ATC level 4 class (~600)

Cohorts generated in five data sources

Merative MarketScan CCAE, MDCD, and MDCR

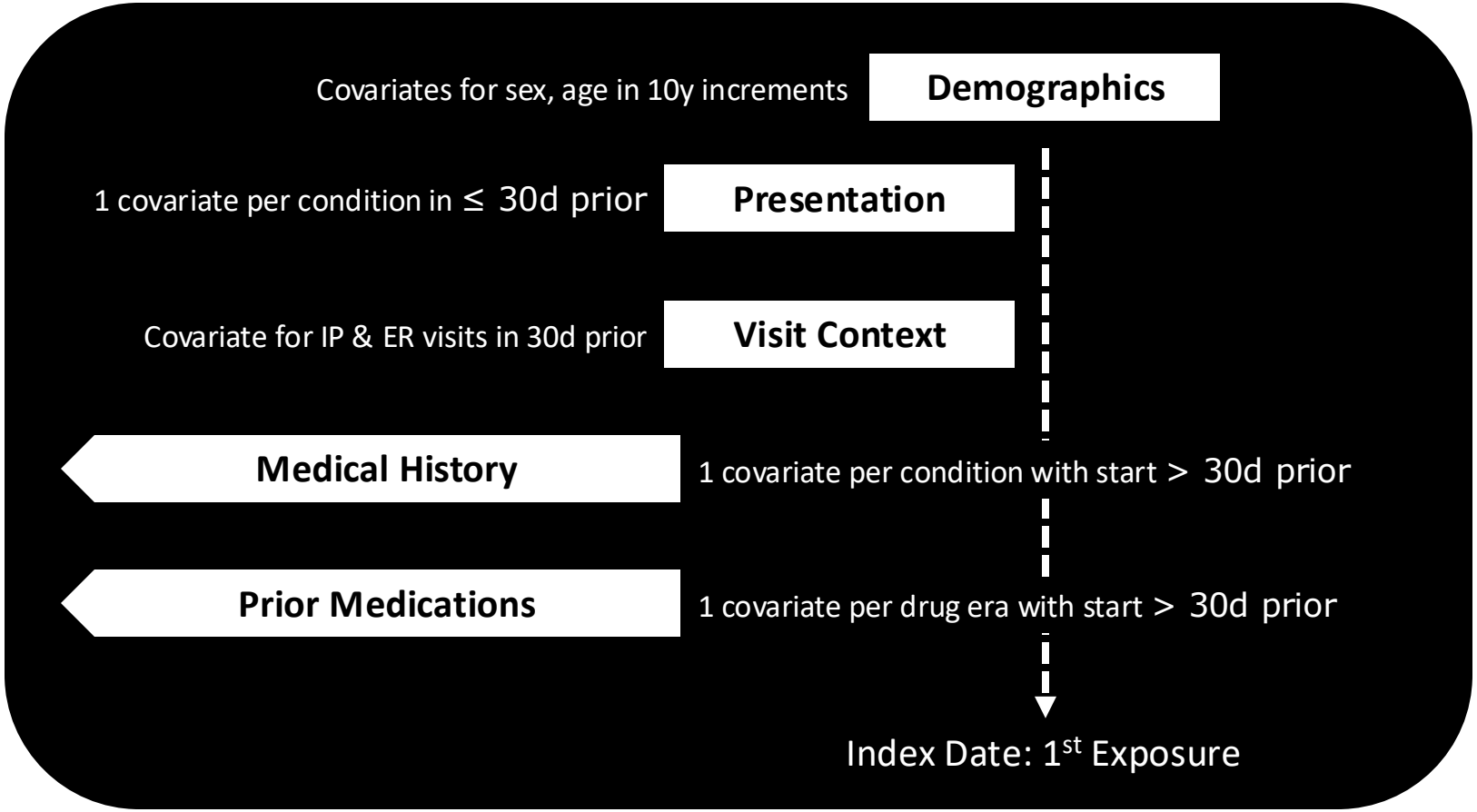
Optum Clinformatics Data Mart

Japan Medical Data Center

Building a cohort similarity pipeline

Data-adaptive covariate data extraction

- SQL extracts aggregated pre-treatment covariate data on all cohorts and stores in separate database table
- Five covariate domains designed to emulate clinical decision-making
- Remove covariates with prevalence < 1%



Building a cohort similarity pipeline

Similarity calculation

- Form all pairs of cohorts with N ≥ 1,000
- Compute **cosine similarity** for each pair, separately by domain
- Final score is the **average of domain-specific cosine similarities**
- For each cohort, all other cohorts can be ranked according to final *cohort similarity score*

The cosine similarity is the cosine of the angle between two vectors, defined as

$$\frac{\boldsymbol{A}\cdot\boldsymbol{B}}{||\boldsymbol{A}||\,||\boldsymbol{B}||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

where ***A*** and ***B*** are vectors of covariate prevalences in the target and comparator cohorts, respectively.
For non-negative numbers like covariate prevalence, cosine similarity will always be between 0 (orthogonal vectors) and 1 (proportional vectors).

Results

Top ranked comparators include clinical alternatives & related treatments

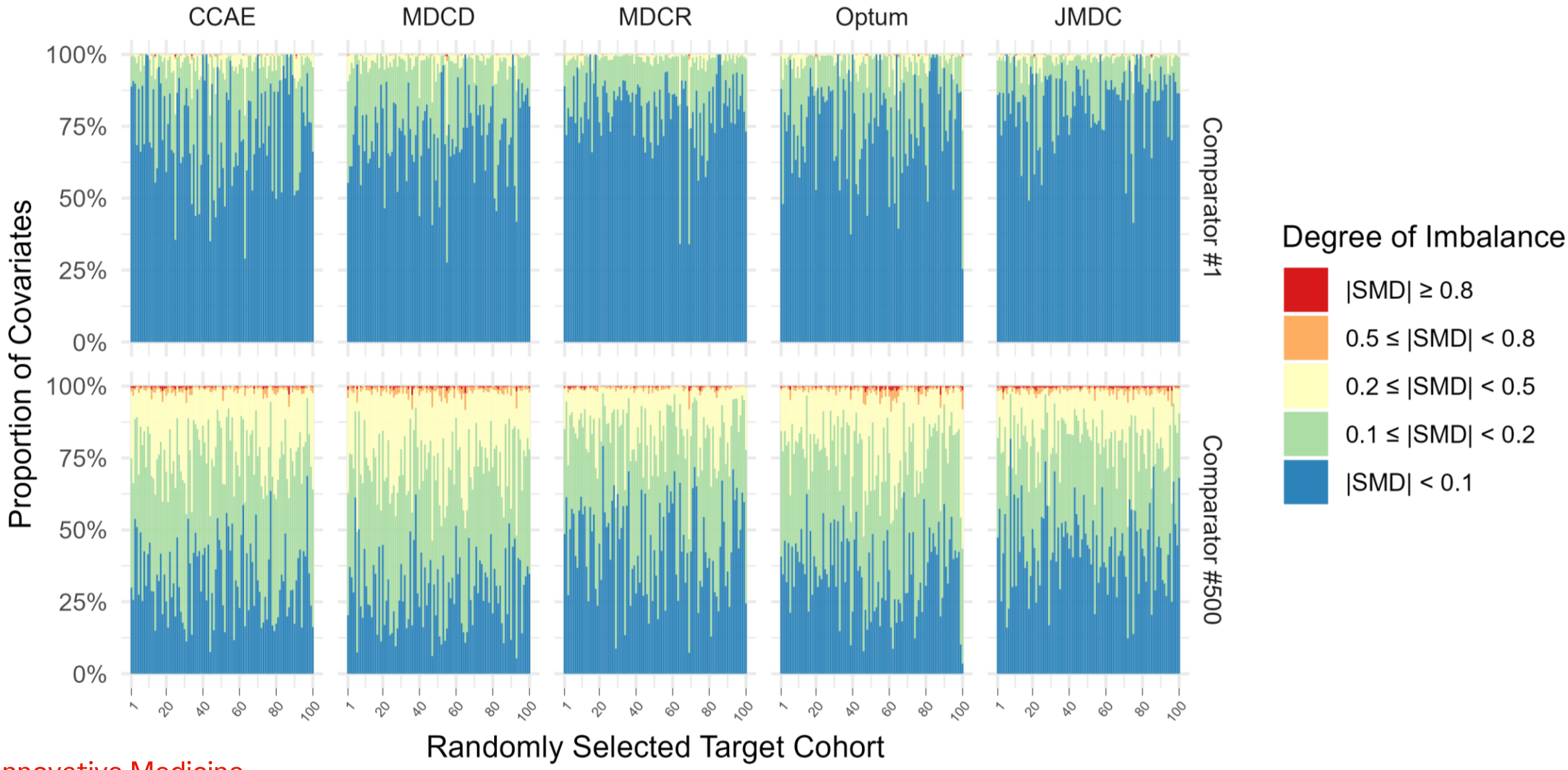
Five most similar comparators by average cohort similarity score for six example drugs*

					Drugs for multiple sclerosis	
Rank	**warfarin**	**apixaban**	**methotrexate**	**adalimumab**	**glatiramer**	**ocrelizumab**
1	rivaroxaban (0.968)	rivaroxaban (0.977)	sulfasalazine (0.978)	certolizumab pegol (0.974)	interferon beta-1a (0.993)	teriflunomide (0.980)
2	digoxin (0.960)	diltiazem (0.955)	leflunomide (0.970)	sulfasalazine (0.960)	interferon beta-1b (0.993)	dimethyl fumarate (0.969)
3	diltiazem (0.959)	amiodarone (0.953)	etanercept (0.966)	infliximab (0.958)	dimethyl fumarate (0.991)	natalizumab (0.967)
4	amiodarone (0.953)	flecainide (0.952)	adalimumab (0.953)	methotrexate (0.953)	fingolimod (0.989)	fingolimod (0.965)
5	carvedilol (0.952)	torsemide (0.947)	certolizumab pegol (0.953)	leflunomide (0.945)	natalizumab (0.988)	dalfampridine (0.967)

* Restricted to comparators appearing in at least two databases

Top-ranked comparators achieve balance on most covariates

Illustration with 100 randomly selected target drugs in each database



Comparators chosen in high-quality observational studies

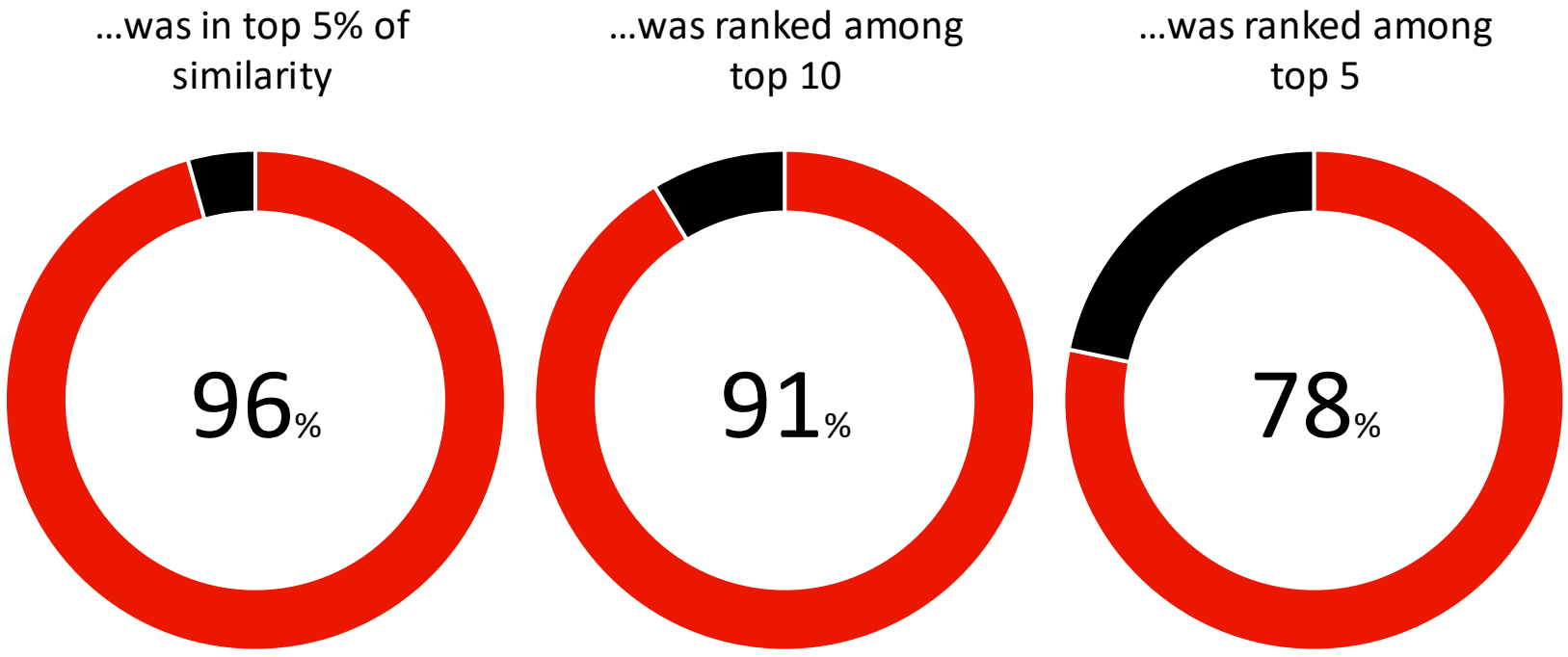
60 papers met search criteria*

24 papers employed NU design

56 comparisons represented

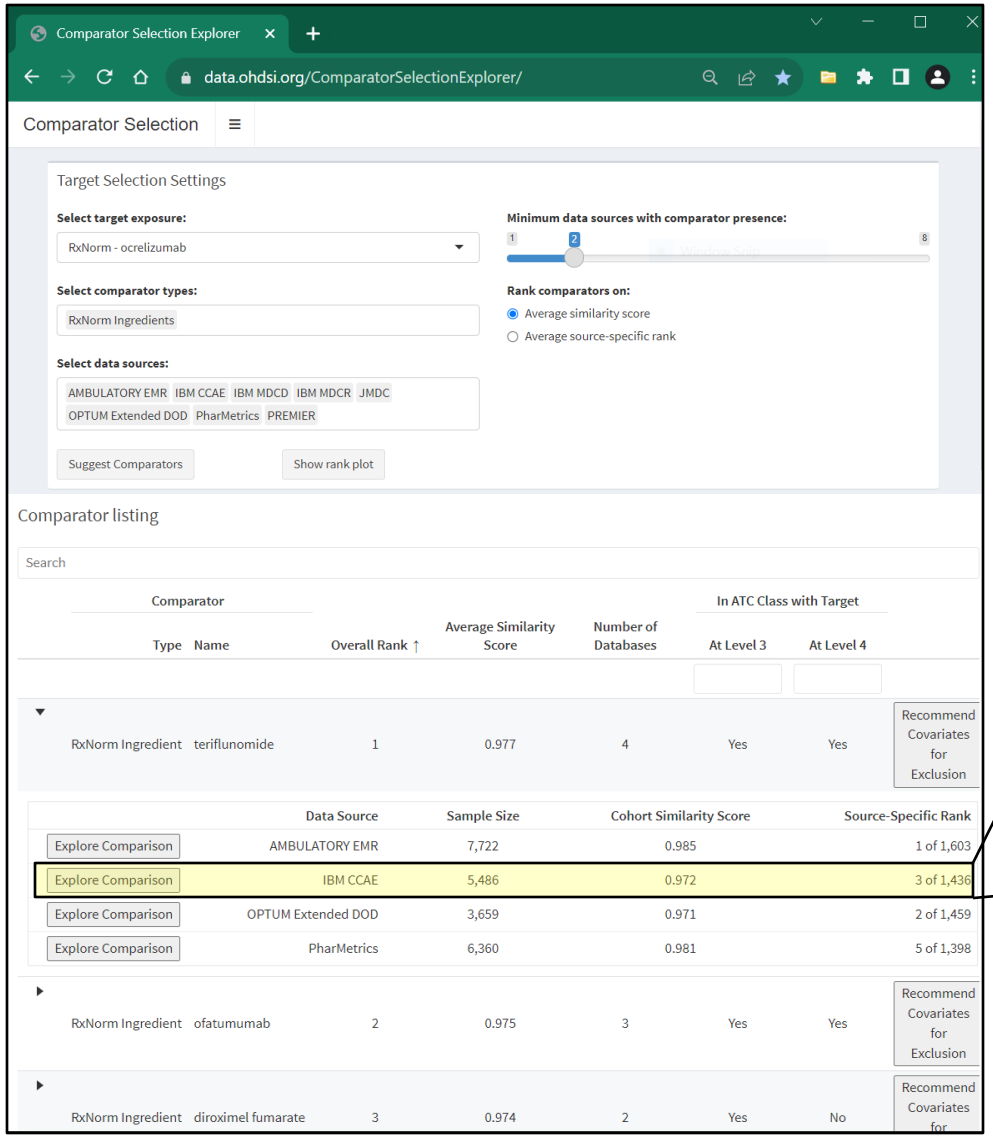
23 comparisons mapped completely

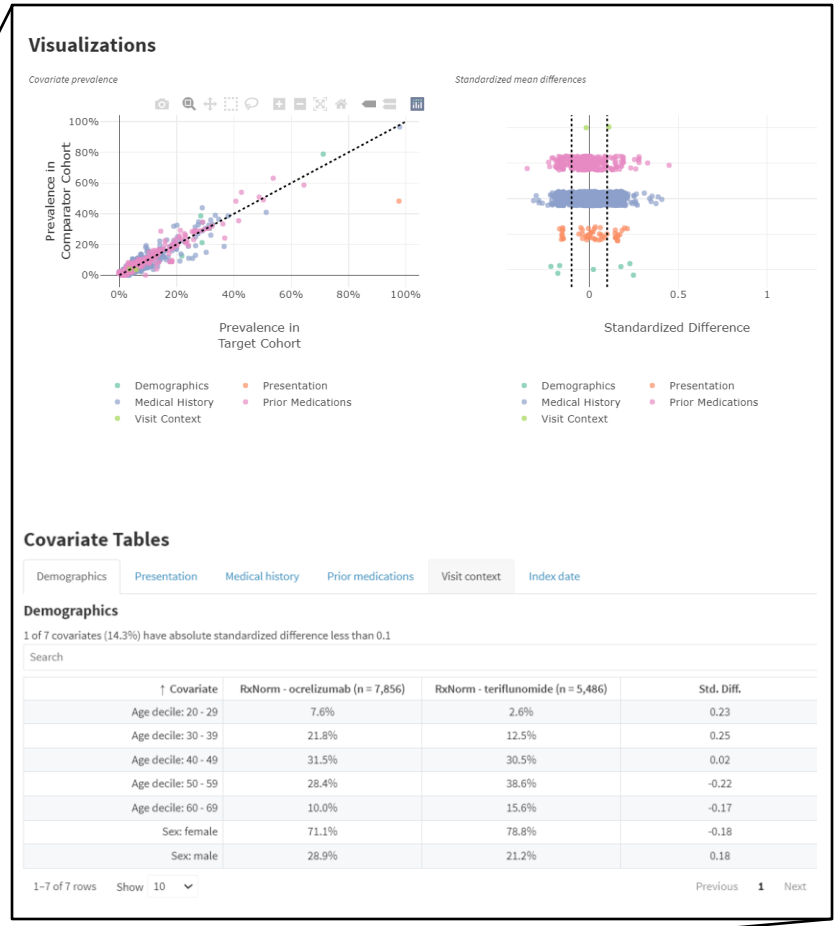
Among the 23 mapped target-comparator pairs, in at least one data source, the chosen comparator …



*Published in high-impact clinical (e.g., JAMA) or pharmacoepidemiological (e.g., PDS) journals in 2021-2023, see supplement for full query details

Exploring comparisons interactively





Scan for Shiny App

Limitations

Of the methodology

- Does not account for
  - Covariate importance
  - Joint distributions
- Similarity ≠ adjustability

Potential mitigations

- Use as screening tool
- Then apply PS methods

Of the recommendations

- Issues may arise for
  - Combination products
  - Drugs with multiple indications
  - Drugs used differently depending on dose, form, route of administration, etc.

Potential mitigations

- Add custom cohort definitions
- Run pipeline in your own data
- Run pipeline nested within an indication

In conclusion

We developed a novel tool for assessing the similarity of new user cohorts at scale and in high-dimensional covariate space

Highly-similar comparators recommended by this tool are

- Clinically relevant
- Align with choices made in high-quality published observational studies
- Already balanced on the majority of covariates with respect to the target

We intend to continue improving this tool and welcome any questions or feedback you may have

Thank you!

Supplemental Material

Disclosures

All authors listed below are employees or contractors of Janssen Research & Development, LLC and may hold stock or stock options in Johnson & Johnson

Justin Bohn

James P. Gilbert

Christopher Knoll

David M. Kern

Patrick B. Ryan

1 Cohort Definition & Generation

New User Cohort Design

Implement for All Drugs

- RxNorm Ingredients (~1,500)
- ATC Level 4 Class (~600)

Generate in 5 Data Sources
Merative MarketScan CCAE, MDCD, and MDCR
Optum Clinformatics Data Mart
Japan Medical Data Center

2 Extraction of Covariate Data

a Extract Data on Covariates in 5 Domains

- Covariates for sex, age in 10-year increments — **Demographics**
- 1 covariate for each condition in 30 days prior — **Presentation**
- Covariates for IP & ER visits in 30 days prior — **Visit Context**
- **Medical History** — 1 covariate for each condition occurring > 30 days prior
- **Prior Medications** — 1 covariate for each drug era with start > 30 days prior

Index Date: 1st Exposure

b Remove covariates with prevalence < 1%

3 Computation of Similarity Scores

a Form all pairs of cohorts with ≥ 1,000 patients

b For each covariate domain within each pair, compute cosine similarity based on vector of covariate prevalences in target cohort (A) and comparator cohort (B)

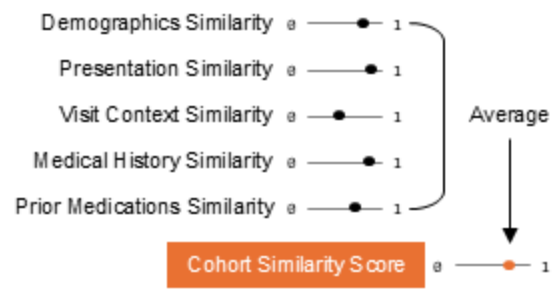
$$A = \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_k \end{bmatrix} \quad B = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_k \end{bmatrix}$$

Cosine Similarity

$$\frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

c

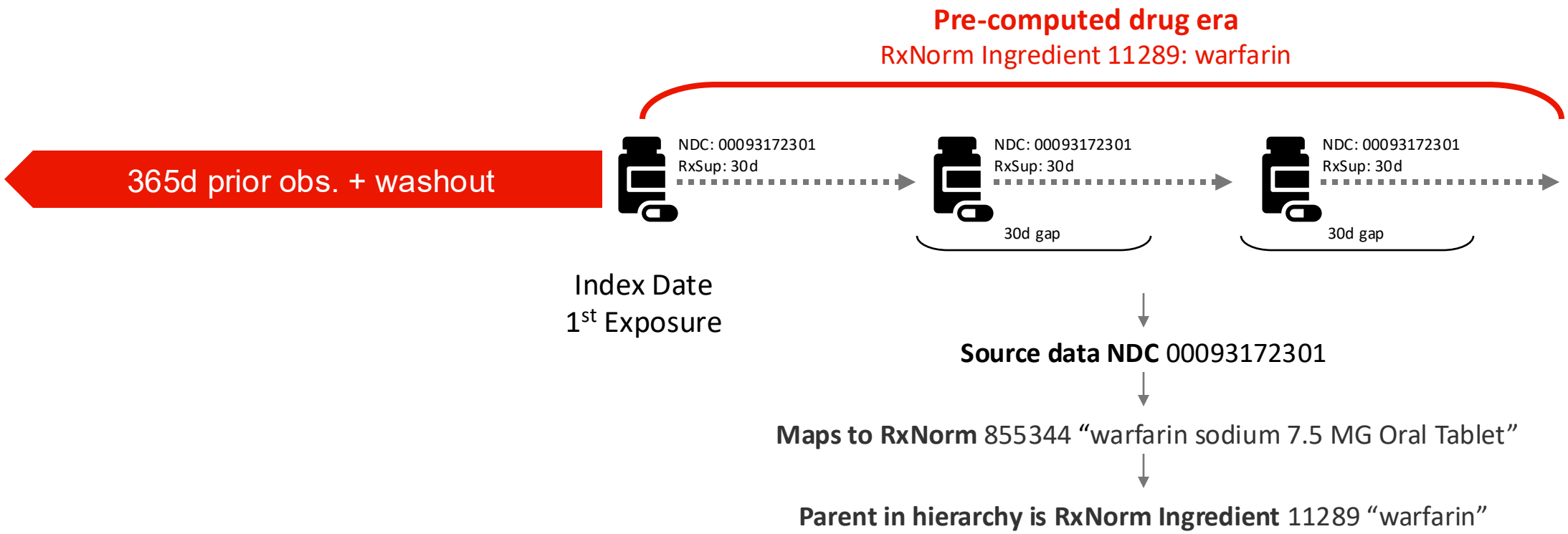
Average Domain-Specific Cosine Similarities



Details on cohort definitions

Repeat for every RxNorm ingredient

New-user cohorts using the OMOP CDM



Summary of cohorts generated, and covariates extracted, by data source and covariate domain

	CCAE	MDCD	MDCR	JMDC	Optum
Total number of cohorts generated					
Overall	2,243	2,158	2,065	1,729	2,273
With at least 1,000 patients	1,359	1,112	988	1,082	1,332
Mean (SD) number of covariates with prevalence ≥ 1% extracted per cohort with ≥ 1,000 patients					
All	1,030.6 (324.65)	1,479.9 (432.59)	1,287.7 (279.28)	934.5 (186.76)	1,422.0 (441.61)
Demographics	7.8 (1.12)	9.6 (1.52)	6.2 (0.57)	8.9 (1.23)	9.3 (1.64)
Medical history	662.3 (234.73)	1,016.4 (319.43)	852.1 (207.44)	336.9 (74.25)	957.3 (309.51)
Presentation	65.4 (49.17)	120.0 (63.71)	88.3 (47.32)	101.5 (36.33)	119.2 (85.42)
Prior medications	293.0 (58.56)	331.9 (67.89)	339.1 (39.66)	486.2 (83.36)	334.3 (64.94)
Visit context	1.8 (0.36)	1.9 (0.25)	1.9 (0.08)	1.0 (0.00)	1.8 (0.34)
Number of pairs of cohorts with ≥ 1,000 patients	922,761	617,716	487,578	584,821	886,446
Mean (SD) similarity score among pairs of cohorts with ≥ 1,000 patients					
Cohort similarity score	0.721 (0.099)	0.774 (0.106)	0.847 (0.060)	0.793 (0.090)	0.743 (0.120)
Demographics	0.852 (0.139)	0.846 (0.128)	0.945 (0.075)	0.857 (0.134)	0.841 (0.147)
Medical history	0.781 (0.137)	0.798 (0.137)	0.913 (0.059)	0.879 (0.092)	0.798 (0.141)
Presentation	0.213 (0.173)	0.401 (0.216)	0.490 (0.190)	0.377 (0.189)	0.349 (0.212)
Prior medications	0.882 (0.085)	0.872 (0.090)	0.931 (0.059)	0.917 (0.071)	0.859 (0.108)
Visit context	0.875 (0.150)	0.955 (0.066)	0.957 (0.069)	1.000 (0.000)	0.869 (0.160)
Number (%) of target cohorts with at least one comparator with cohort similarity score of:					
At least 0.900	1,294 (95.2%)	1,094 (98.4%)	981 (99.3%)	1,067 (98.6%)	1,294 (97.1%)
At least 0.950	1,087 (80.0%)	961 (86.4%)	924 (93.5%)	1,024 (94.6%)	1,118 (83.9%)

Drugs with known relationships are rated as more similar

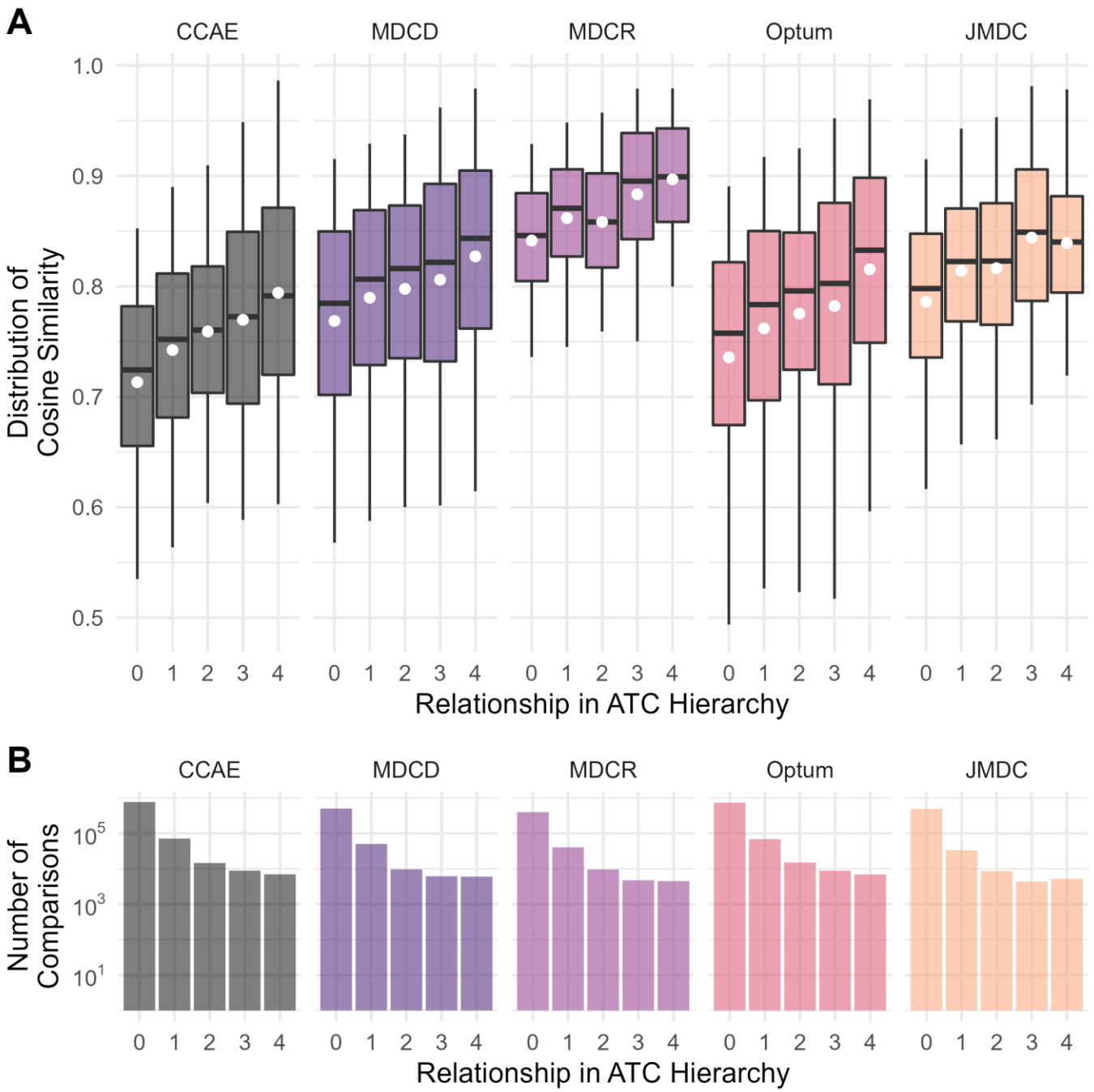
Related at level 3

Level 1: *B* Blood and blood forming organs
Level 2: *B01* Antithrombotic agents
Level 3: *B01A* Antithrombotic agents
Level 4: *B01AA* Vitamin K antagonists
Level 5: *B01AA03* warfarin

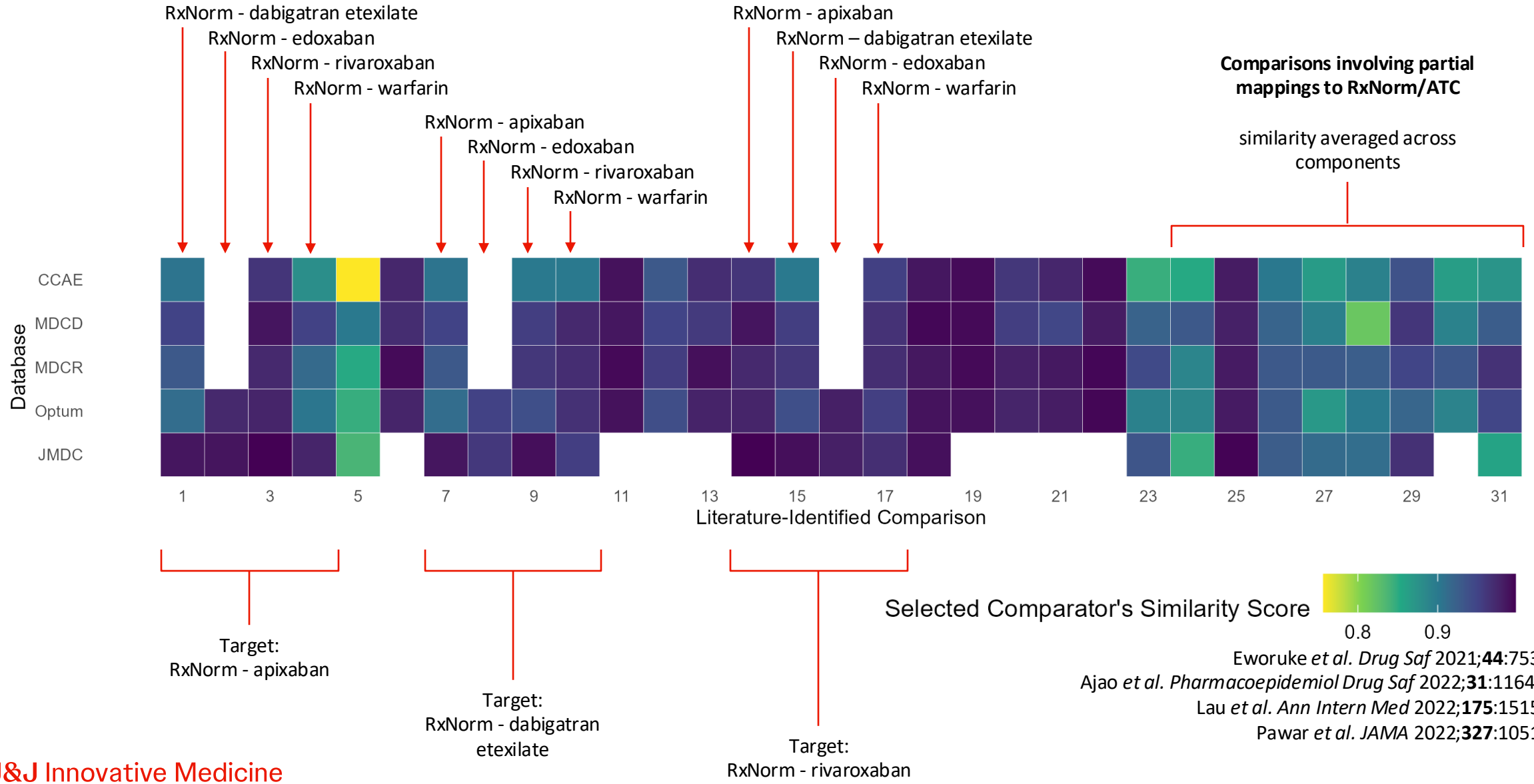
Level 1: *B* Blood and blood forming organs
Level 2: *B01* Antithrombotic agents
Level 3: *B01A* Antithrombotic agents
Level 4: *B01AF* Direct factor Xa inhibitors
Level 5: *B01AF02* apixaban

Unrelated (i.e., "0")

Level 1: *N* Nervous system
Level 2: *N05* Pyscholeptics
Level 3: *N05A* Antipsychotics
Level 4: *N05AH* Diazepines, oxazepines, thiazepines and oxepines
Level 5: *N05AH02* clozapine



Comparators chosen in high-quality publications are highly similar



Eworuke *et al. Drug Saf* 2021;**44**:753–63
Ajao *et al. Pharmacoepidemiol Drug Saf* 2022;**31**:1164–73.
Lau *et al. Ann Intern Med* 2022;**175**:1515–24
Pawar *et al. JAMA* 2022;**327**:1051–60

Search criteria for the literature review

Publication year: 2021-2023

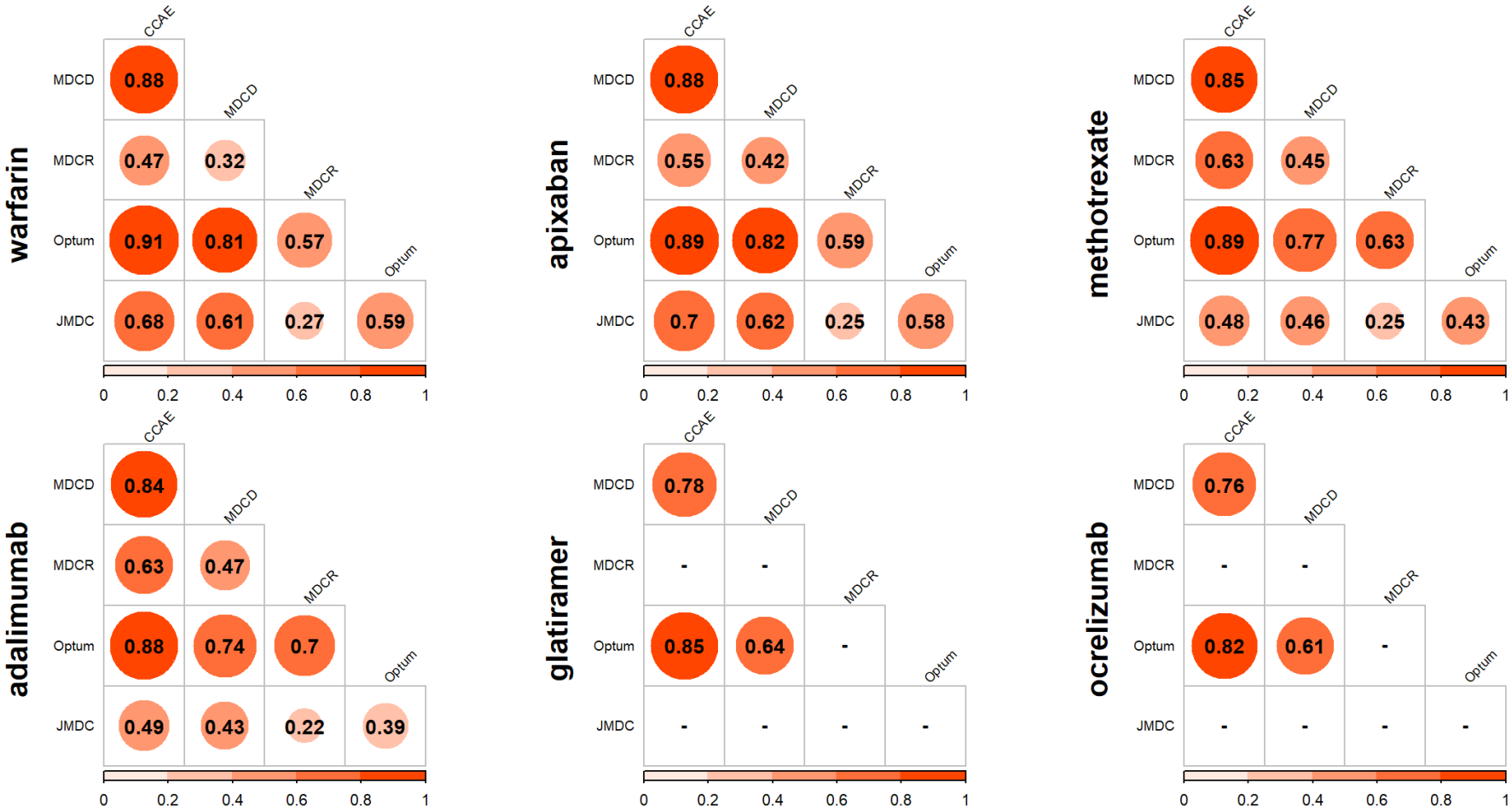
Journals

- Science, Nature, The New England Journal of Medicine, JAMA, JAMA Internal Medicine, JAMA Surgery, JAMA Psychiatry, JAMA Cardiology, JAMA Pediatrics, JAMA Neurology, The Lancet, BMJ, Diabetes Care, Hypertension, Journal of the American College of Cardiology, Pharmacoepidemiology and drug safety, Drug safety, Annals of internal medicine, The Lancet. Digital health, The Lancet. Infectious diseases, The lancet. Diabetes endocrinology, The Lancet. Neurology, The Lancet. Oncology, The Lancet. Respiratory medicine, The lancet. Psychiatry, The Lancet. Global health

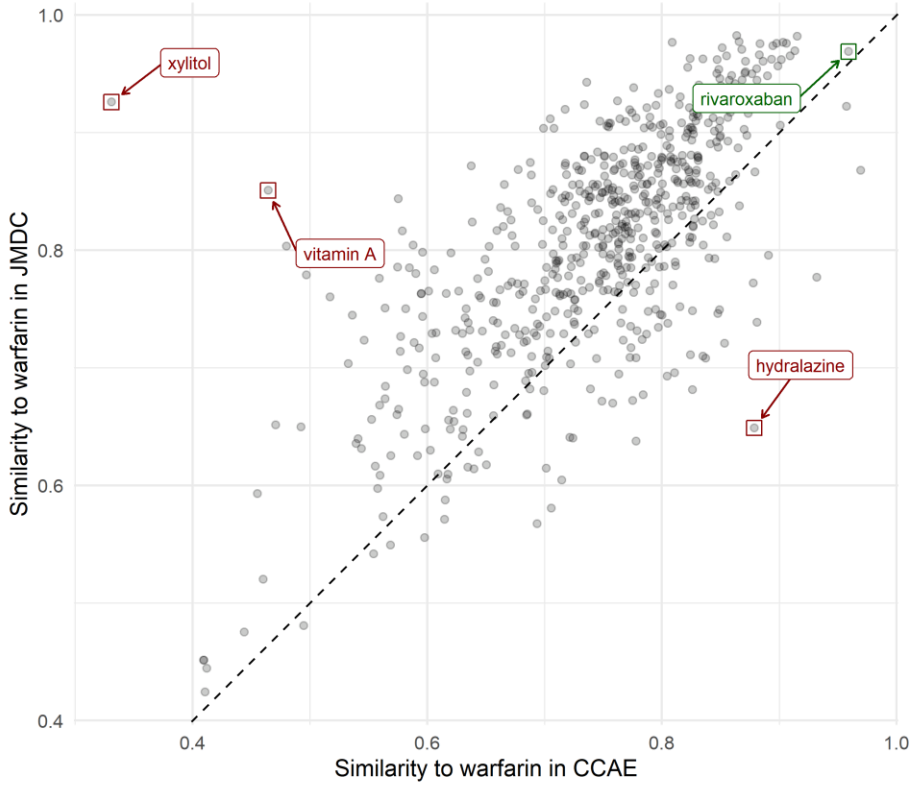
Required terms:

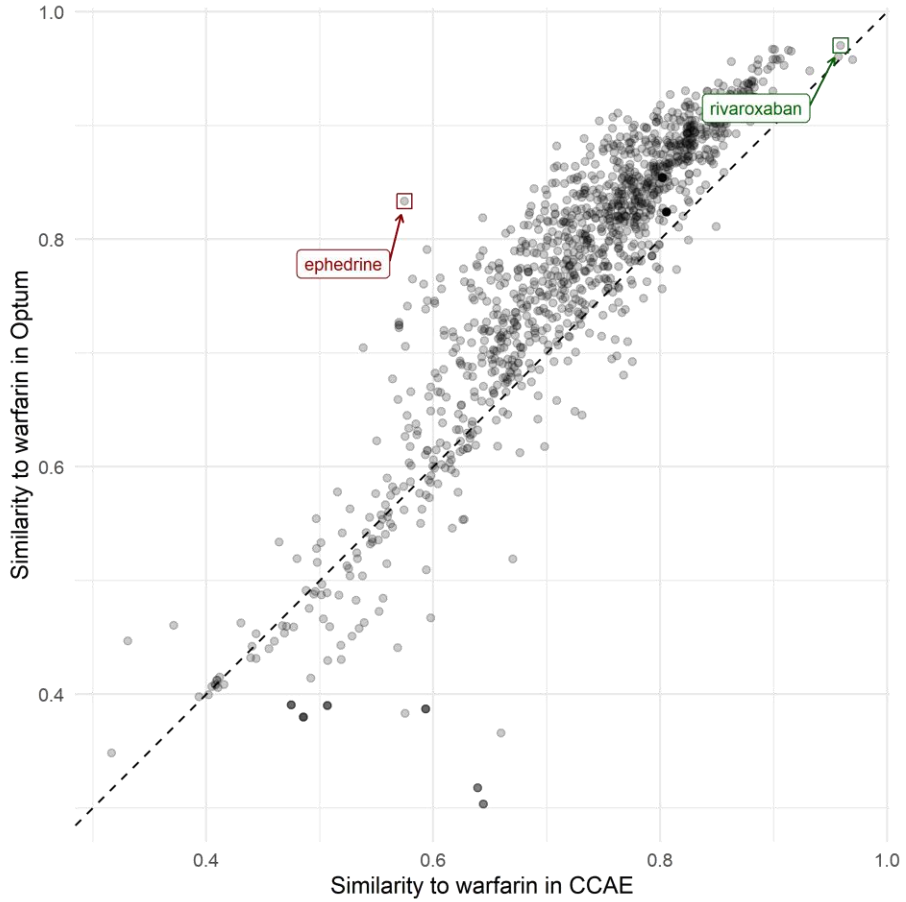
- ("marketscan" OR "truven" OR "optum" OR "clinformatics" OR "iqvia" OR "pharmetrics" OR "administrative claims" OR "insurance claims" OR "healthcore" OR "aetna" OR "cigna" OR "humana" OR "sentinel") AND ("cohort" AND ("comparator" OR "compared" OR "comparative" OR "propensity score" OR "versus")) AND ("risk" OR "effect" OR "causal" OR "hazard" OR "exposure" OR "RR" OR "HR" OR "IRR" OR "rate ratio") NOT ("cross sectional") NOT ("Randomized Controlled Trial"[Publication Type] OR "Meta-Analysis"[Publication Type])

Cross-database correlation in similarity



Cross-database correlation in similarity

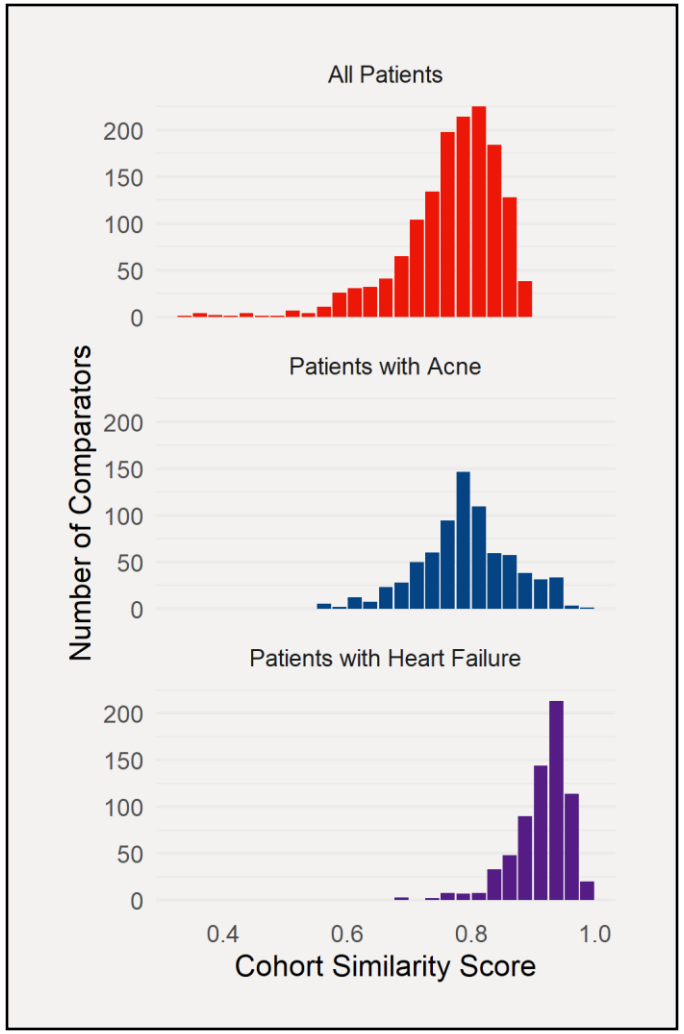




Restriction based on indication can meaningfully change recommendations

Five most similar comparators for target **spironolactone** within different patient populations

Rank	All Patients 1,476 Comparators	Patients with Acne 63% Reduction in Target 758 Comparators	Patients with Heart Failure 89% Reduction in Target 690 Comparators
1	furosemide (0.897)	azelate (0.978)	carvedilol (0.992)
2	clindamycin (0.895)	dapsone (0.969)	bisoprolol (0.990)
3	Lactobacillus acidophilus (0.895)	triamcinolone (0.956)	torsemide (0.988)
4	ethacrynate (0.894)	levonorgestrel (0.952)	furosemide (0.987)
5	enoxaparin (0.892)	etonogestrel (0.947)	lisinopril (0.986)



Data from Merative MarketScan CCAE

For acne and heart failure analyses, all target and comparator cohorts restricted to patients with the relevant condition recorded within the 365 days prior to index

A first attempt at empirical comparator identification

Identify cohorts with ≥ 1,000 patients

- First use of each RxNorm ingredient with 365 days prior observation

Randomly sample 1,000 patients from each cohort to reduce computation time

Use FeatureExtraction to extract features

Form all possible pairs of cohorts

Restrict to only pairs involving six drugs of interest to reduce computation time

Use FeatureExtraction to compute standardized differences for all features in a given pair

Average absolute standardized differences to get similarity score

Rank according to average absolute standardized difference

Some reading on comparator selection

The active comparator new user design

- Lund JL, Richardson DB, Stürmer T. The Active Comparator, New User Study Design in Pharmacoepidemiology: Historical Foundations and Contemporary Application. *Curr Epidemiol Rep* 2015;**2**:221–8. doi:10.1007/s40471-015-0053-5

Comparator selection recommendations

- Setoguchi S, Gerhard T. Chapter 5: Comparator Selection. In: Velentgas P, Dreyer NA, Nourjah, P, *et al.*, eds. *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide*. Rockville (MD): : Agency for Healthcare Research and Quality (US)

Negative control and null comparator methodology

- Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology* 2010;**21**:383–8. doi:10.1097/EDE.0b013e3181d61eeb
- Dusetzina SB, Brookhart MA, Maciejewski ML. Control Outcomes and Exposures for Improving Internal Validity of Nonrandomized Studies. *Health Serv Res* 2015;**50**:1432–51. doi:10.1111/1475-6773.12279
- Huitfeldt A, Hernan MA, Kalager M, *et al.* Comparative Effectiveness Research Using Observational Data: Active Comparators to Emulate Target Trials with Inactive Comparators. *EGEMs Gener Evid Methods Improve Patient Outcomes* 2016;**4**:20. doi:10.13063/2327-9214.1234

Negative control and null comparator example studies

- Setoguchi S, Glynn RJ, Avorn J, *et al.* Statins and the Risk of Lung, Breast, and Colorectal Cancer in the Elderly. *Circulation* 2007;**115**:27–33. doi:10.1161/CIRCULATIONAHA.106.650176
- Solomon DH, Avorn J, Stürmer T, *et al.* Cardiovascular outcomes in new users of coxibs and nonsteroidal antiinflammatory drugs: High-risk subgroups and time course of risk. *Arthritis Rheum* 2006;**54**:1378–89. doi:10.1002/art.21887

A note on defining covariate imbalance (1)

STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:3083–3107
Published online 15 September 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3697

Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples

Peter C. Austin[1, 2, 3, *, †]

residual imbalance between treated and untreated subjects in the matched sample. While there is no clear consensus on this issue, some researchers have proposed that a standardized difference of 0.1 (10 per cent) denotes meaningful imbalance in the baseline covariate [23]. It is likely that the

23. Normand SLT, Landrum MB, Guadagnoli E, Ayanian JZ, Ryan TJ, Cleary PD, McNeil BJ. Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of Clinical Epidemiology* 2001; **54**:387–398.

A note on defining covariate imbalance (2)

ELSEVIER

Journal of Clinical Epidemiology 54 (2001) 387–398

Journal of Clinical Epidemiology

Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores

Sharon-Lise T. Normand[a,b,*], Mary Beth Landrum[a], Edward Guadagnoli[a], John Z. Ayanian[a,e], Thomas J. Ryan[d], Paul D. Cleary[a,c], Barbara J. McNeil[a,f]

[a]*Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA, 02115-5899, USA*
[b]*Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA*
[c]*Department of Social Medicine, Harvard School of Public Health, Boston, MA, USA*
[d]*Boston University School of Medicine, Boston, MA, USA*
[e]*Division of General Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA*
[f]*Department of Radiology, Brigham and Women's Hospital, Boston, MA, USA*

Received 4 April 2000; received in revised form 8 July 2000; accepted 8 August 2000

ances. Small (<10%) absolute values of d_i support the assumption of balance [25] between treatment groups.

[25] Cohen J. Statistical power analysis for the behavioral sciences. Toronto: Academic Press, Inc., 1977 [chapter 2].

A note on defining covariate imbalance (3)

Statistical Power Analysis for the Behavioral Sciences

Second Edition

Jacob Cohen
Department of Psychology
New York University
New York, New York

LAWRENCE ERLBAUM ASSOCIATES, PUBLISHE

2. When **d** = .1 as in the above example, the distribution of the population with the larger mean, B, is almost superimposed on A, but with some slight excess, i.e., some nonoverlap. U_1 here equals 7.7%, that is, 7.7% of

2.2.3 "SMALL," "MEDIUM," AND "LARGE" d VALUES.

SMALL EFFECT SIZE: **d** = .2. In new areas of research inquiry, effect sizes are likely to be small (when they are not zero!). This is because the

MEDIUM EFFECT SIZE: **d** = .5. A medium effect size is conceived as one large enough to be visible to the naked eye. That is, in the course of

LARGE EFFECT SIZE: **d** = .8. When our two populations are so separated as to make **d** = .8, almost half (U_1 = 47.4%) of their areas are not overlapped. U_2 = 65.5%, i.e., the highest 65.5% of the B population exceeds